DOCUMENT RESUME

ED 223 135 HE 015 519

AUTHOR Williams, Benne D.
TITLE Teacher and Course Evaluations That Really Discriminate.
PUB DATE 4 Apr 82
NOTE 15p.; Paper presented at the Annual Meeting of the Rocky Mountain Psychological Association (April 28-May 1, 1982).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Faculty; *College Instruction; *Course Evaluation; *Evaluation Methods; Higher Education; Measurement Techniques; *Rating Scales; Student Attitudes; *Student Evaluation of Teacher Performance; *Teacher Effectiveness; Test Validity

ABSTRACT

A modified rank order procedure was developed and tested to increase the sensitivity of student evaluation instruments. Students in 14 psychology classes at Weber State College were asked to think of all the courses they have taken at the college level and to determine which one was the best and which course or instructor was the worst. These courses were criteria against which the present course was evaluated. Students were asked to rank the course or instructor being evaluated on a seven-point scale, the Endeavor Rating Scale. Students were randomly divided into two groups in each class to compare the traditional evaluation instrument used at the college and other institutions and the rank comparison scale. Findings suggest that the rank comparison scale method was more sensitive to differences between instructors than was the traditional seven-point rating scale. The procedure did not appear to detract from the validity of the instrument when measured against supervisory ratings and the nominations made in the instrument for best and worst instructor/course. The rank comparison scale procedure can be used with any number of items from an item pool selected either by the faculty, students, or school administrators. Recommendations include the following: the item pool to be used with the rank comparison method should be carefully selected in order to measure precisely the elements of the instructional procedure and/or instructor characteristics that are relevant to a specific course/instructor; and all of the items should carry operational definitions of the characteristics being measured. (SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED223135

Teacher and Course Evaluations That Really Discriminate

by

Benne D. Williams

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Benne D. Williams

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Psychology Department
Weber State College
Ogden, Utah
1982

HE 015 519

Presented at RMPA, April 4, 1982

Since the work of G. C. Brandenburg and H. H. Rimmers (2) in 1927, ratings of public school and college instructors have been the subject of extensive experimentation and research. The variation with respect to the characteristics being rated and the particular format being used to obtain these ratings has been enormous. For the most part, the use of the ratings has remained relatively stable. Most often they are used for feedback to teachers and for administrative purposes such as tenure, salary, and promotional considerations. Despite the volumes which have been written regarding student ratings, little has been done to develop instruments which were sensitive enough to clearly discriminate between teachers with varying levels of acclaimed effectiveness.

Most of the research reported by such reviewers as Costen, Greenough, Menges (5), and Aleamoni (1) are replete with research studies on reliability, validity, and the usefulness of student ratings of teaching. The most consistent result of students rating their teachers and courses have been that almost all faculty are rated above average with very little difference between those rated high and those rated at the lowest value level. In my own experience as chairman of a department of psychology for six years, there was not one instance in which any instructor in a ten-person department received student ratings which were below the mid-point of the scale being used. Likewise, there were very few instances when there was a significant difference between the ratings of the highest person and those of the lowest person.

With this situation prevalent throughout the country, it is easy to see why improvement of instruction at the college level has been difficult

to motivate, let alone achieve. College faculty members generally see themselves as being experts in pedagogy even though most of them have never received formal instruction in teacher training. It is likely that these deductions, regarding their expertise as teachers, have been drawn from student ratings which have not been sensitive enough to discriminate between effective and ineffective teachers. One hears continual complaints about the ineffectiveness of college instructors, but one seldom sees any evidence of this ineffectiveness in student ratings. Thus, spuriously high ratings given to college teachers by their students have been counterproductive both as a measure of teaching effectiveness and as a motivation for the improvement of instruction.

Development of the Method

This study concerns itself with the development of an evaluation procedure which is intended to increase the sensitivity of any student evaluation instrument and render more realistic ratings. Using the idea of Herbert W. March (9) who asked his graduating seniors to nominate instructors who were "most outstanding" or "least outstanding," a modified rank order procedure was developed to increase the sensitivity of the items on any rating scale. Students were asked to think of all the courses they have taken at the college level and determine which one of them was the best and which they would rate as the worst course or instructor. They were asked to write the names of these courses and instructors on the rating sheet. These courses were to serve as the criteria against which the present course would be evaluated. Next the students were asked to rank the course or instructor being evaluated on

each of the items in the evaluation form according to the following ranks:

Rank	Description
1	Ranks BELOW THE WORST course/instructor I have ever had, a new worst course.
2	Ranks EQUAL TO THE WORST course/instructor I have ever had.
3	Ranks ABOVE THE WORST course/instructor, but below the average course/instructor I have ever had.
4	Ranks EQUALLY BETWEEN WORST AND BEST course/instructor I have ever had.
5	Ranks just BELOW THE BEST course/instructor I have ever had.
6	Ranks EQUAL TO THE BEST course/instructor I have ever had.
7	Ranks ABOVE THE BEST course/instructor I have ever had, establishing a new BEST COURSE.

While this rating format may be utilized with any number of different item pools, for this research the seven-item Endeavor Rating Scale was used. The format has since been used with a thirty-item rating scale and a ten-item rating scale to determine if the amount of time necessary to complete the rating was overly laborious and counterproductive. An examination of the use of different item pools used with this procedure reveals little difference in the students' responses and attitudes. Students were willing to respond conscientiously regardless of the length of the item pool. They reported that they were required to exert equal consideration, effort and thought on each item on the instrument regardless of its ordinal location.

Procedure

Students in fourteen psychology classes as Weber State College were randomly divided into two groups in each class. One group received the traditional evaluation instrument used at Weber State College. The other group received the Rank Comparison Scale under investigation in

this study. Written instructions for the use of each instrument were used so that each student would not be given an improper mental set by overhearing instructions intended for someone else. Students made their ratings anonymously while the teacher was absent from the room, and any questions were answered on an individual basis by the trained proctor who was from the secretarial pool of the department. The students were permitted as much time as was necessary to complete the evaluation form and were encouraged to make written comments as well as marking the answer sheet objectively.

The results were tabulated by a computer program which also computed averages and standard deviations for each of the instructors under each of the conditions. Computations of the standard error of measurement and the standard difference between the means with T scores were also computed. Significance values were determined using a statistical table of the percent of total area under the normal curve.

Results

Results of the traditional evaluation instrument are expressed using a seven-point rating scale from "strongly disagree" to "strongly agree." Table 1 shows the results of student ratings for the 14 faculty

- - - - - - - - - - - - - - - - - -

INSERT TABLE 1 ABOUT HERE

- - - - - - - - - - - - - - - - - -

members evaluated with the resulting means and standard deviations. The range on the evaluations was from 6.25 to 5.46, a difference of only .79. The departmental mean was 5.81. Not one of the 14 professors obtained a

significant deviation from the departmental mean. The difference between the means of instructors A and M was significant at the .01 level, and the difference between the means of professors A and L, as well as B and L, were significant at the .05 level. None of the other mean differences were significant. Not one instructor scored below the 3.5 midpoint of the scale. In fact, the lowest ranking professor received a spurious 78 percent of the total points possible.

The data for the new rating scale is shown in Table 2 with mean

- - - - - - - - - - - - - - - - - - - -

INSERT TABLE 2 ABOUT HERE

- - - - - - - - - - - - - - - - - - - -

scores and standard deviations. It will be noted that the range for the Rank Comparison Scale was from 6.1 to 3.3, a difference of 2.8. This is a range three times larger than the traditional rating scale. The departmental mean was 4.8, one full score lower than that achieved with the traditional instrument. One instructor scored below the midpoint on the scale, a first in the experience of this author. Table 3 shows

- - - - - - - - - - - - - - - - - - - -

INSERT TABLE 3 ABOUT HERE

- - - - - - - - - - - - - - - - - - - -

the significant differences which existed between the various instructors at the .01 and .05 levels. Contrasts between the ability of the traditional rating scale and that of the newer Rank Comparison Scale to produce significant differences between the means of the instructors rated is encouraging and demonstrates some promise as an instrument

sensitive to differences often observed by students but seldom measured. While the traditional rating scale surprisingly yielded one mean difference significant at the .01 level, the Rank Comparison Scale yielded 20 differences significant at the .01 level and 10 at the .05 level.

The validity of the Rank Comparison Scale was tested by computing a correlation between the scale and the supervisors' ratings of the 14 professors. The correlation was .74.

The best and worst nominations were converted to standard scores and then placed into rank order. This made it possible to compare the ratings given to each of the professors on the basis of being nominated best or worst and the average ratings given to those professors for each of the items on the rating scale. Correlation between the best and worst nominations and the items on the Rank Comparison Scale was .85. This might be considered a measure of internal consistency and suggests a measure of reliability for the instrument.

Conclusions

The Rank Comparison Scale method appears to be much more sensitive to differences that exist between various instructors than the traditional seven-point rating scales currently in use in many institutions. The procedure does not seem to detract from the validity of the instrument when measured against supervisory ratings and the nominations made in the instrument itself for best and worst instructor/course. The Rank Comparison Scale procedure can be used with any number of items from an item pool selected either by the faculty, students, or by school administrators.

It was observed that some of the items on the instrument used in this study were not as relevant to some classes as they were to others. This generally resulted in a high standard deviation which made it more difficult to obtain significant differences between the means of the various professors. The use of the rank comparison technique produced a greater range between the high and low scores, a larger number of statistically significant mean differences, and increased the sensitivity and discrimination of each of the items on the evaluation instrument. It was also noticed that response set (a tendency to score all of the items of a particular pool at the same value rating) was almost entirely eliminated by the use of the rank comparison method. Thus it would appear that this procedure would yield more realistic evaluation of the effectiveness of the instructor and a more detailed and discriminating analysis of the various types of instructional activities measured by the various items of the evaluting instrument. Eliminating spurious evaluations and providing more discriminating analysis of the performance of the instructor may lead to a conclusion that improvement in instruction is not only advisable but necessary for some instructors. Hopefully, this would provide the motivation for modification of their instructional philosophy and/or behavior.

Recommendations

Experience with this study and results of the study itself suggest five additional recommendations:

1. The item pool to be used with the rank comparison method should be carefully selected so as to measure precisely

those elements of the instructional procedure and/or instructor characteristics which are relevant to a specific course or instructor.

2. The item pool should be stated in such a way as to increase the possibility of a higher inter-rater reliability which in turn would lower the standard deviation for the scores on each of the items. This in turn would result in the possibility of more significant differences between the means of various instructors. This can be done by insuring that each rater reads and understands the instrument exactly the same way as all other raters.
3. Items rating the characteristics of the course should be separated from those items rating the characteristics of the instructor.
4. All of the items should carry operational definitions of the characteristics being measured, and the items should be written in the student vocabulary.
5. The more that faculty members can be involved in selection of items to measure their own effectiveness, the more likely it is that the measurement will be considered valid to that instructor and the greater the possibility it may motivate some instructional changes. Frequently, a discussion of the items to be rated with the instructor's supervisor prior to the evaluation and again after results have been returned can produce a situation in which profitable goal-setting

interviews can be conducted which will not only specify the behavior change necessary but the economic and human resources available to assist the instructor in making these changes. This goal-setting session could then become the objective basis for the evaluation of that professor's teaching effectiveness on future occasions.

TABLE 1

Means and Standard Deviations From Student Ratings
Using Traditional Instrument

Instructor	Mean Rating	Standard Deviation
A	6.25	.71
B	6.24	.60
C	6.17	.68
D	6.16	.61
E	6.13	.70
F	6.06	.79
G	5.84	.88
H	5.74	.93
I	5.73	.87
J	5.63	1.03
K	5.60	.98
L	5.51	.99
M	5.49	.97
N	5.47	.99
Dept. Mean	5.81	.85

Range 5.47 to 6.25 = .79

TABLE 2

Means and Standard Deviations From Student Ratings
Using the "Rank Comparison Scale"

Instructor	Mean Rating	Standard Deviation
A	6.10	.81
C	5.90	.52
B	5.70	.97
D	5.65	1.73
E	5.60	1.32
F	5.40	.92
G	4.90	1.69
H	4.90	1.49
K	4.80	1.60
J	4.60	1.29
N	4.50	1.30
L	3.90	1.87
I	3.80	1.31
M	3.30	1.96
Dept. Mean	4.50	.87

Range 3.30 to 6.10 = 2.8

TABLE 3

Significant Mean Differences Beteeen Instructors
With T Values and Significance

Instructor	Mean	Instructor	Mean	Difference	T Value	Sig.
A	6.1	F	5.4	.7	3.07	.01
		G	4.9	1.2	2.80	.01
		H	4.9	1.2	2.31	.05
		I	4.8	1.3	2.21	.05
		J	4.6	1.5	4.95	.01
		K	4.5	1.6	4.65	.01
		L	3.9	2.2	3.91	.01
		M	3.8	2.3	4.13	.01
		N	3.3	2.8	2.82	.01
B	5.9	G	4.9	1.0	2.08	.05
		H	4.9	1.0	1.79	___ *
		I	4.8	1.1	1.77	___ *
		J	4.6	1.3	3.51	.01
		K	4.5	1.4	4.24	.01
		L	3.9	2.0	3.33	.01
		M	3.8	2.1	3.56	.01
		N	3.3	2.6	2.58	.01
C	5.7	J	4.6	1.1	3.67	.01
		K	4.5	1.2	3.43	.01
		L	3.9	1.8	3.21	.01
		M	3.8	1.9	3.39	.01
		N	3.3	2.4	2.42	.05
D	5.6	L	3.9	1.7	2.21	.05
		M	3.8	1.8	2.35	.05
		N	3.3	2.3	2.05	.05
E	5.65	L	3.9	1.75	2.28	.05
		M	3.8	1.85	2.73	.01
		N	3.3	2.35	2.19	.05
F	5.4	J	4.6	.8	2.61	.01
		K	4.5	.9	2.59	.01
		L	3.9	1.5	2.66	.01
		M	3.8	1.6	2.84	.01
		N	3.3	2.1	2.11	.05

*Not significant.

SELECTED REFERENCES

1. Aleamoni, L. M., and Hexner, P. Z. A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluations. Instructional Science, 1980, 9:67-84.

2. Brandenburg, G. C., and Remmers, H. H. The Purdue Rating Scale for Instructors. Educational Administration and Supervision, 1927, 13:399,406.

3. Borgatta, E. F. Student ratings of faculty. American Association of University Professors Bulletin, 1970, 56:6-7.

4. Clark, K. E. and Keller, R. J. Student ratings of college teaching, in Eckert, R. A. (Ed.) A University Looks At Its Progarm. Minneapolis, University of Minnesota Press, 1954.

5. Costin, F. W., Greenough, W. T., and Menges, R. J. Student ratings of college teaching; Reliability, validity, and usefulness. Review of Educational Research, 1971, 41:511-35.

6. DOYLE, K. O., and Whitely, S. E. Student ratings as a criteria for effective teaching. American Educational Research Journal, 1974, 11: 259-74.

7. Frey, P. W. A Two-dimensional analysis of student ratings of instruction. Research in Higher Education, 1978, 9:69-91.

8. Lovell, G. D., and Haner, C. F. Forced-choice applied to college faculty rating. Educational and Psychological Measurement, 1955, 15:291-304.

9. Marsh, H. W. The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. American Educational Research Journal, 1977, 14:441-47.

10. Nillmann, J. (Ed.) Handbook of Teacher Evaluation. Beverly Hills, Sage, 1981.

11. Reardon, M., and Waters, L. K. Leniency and halo in student ratings of college instructors: A comparison of three rating procedures with implications for scale validity. Educational and Psychological Measurement, 1979, 39:159-162.

12. Sullivan, A. M., and Skanes, G. R. Validity of Student evaluation of teaching and the characteristics of successful instructors. Journal of Educational Psychology, 1974, 66: 584-590.